Event-Driven Deep Neural Network Hardware System for Sensor Fusion

Kiselev, Ilya ; Neil, Daniel ; Liu, Shih-Chii

Abstract: This paper presents a real-time multi-modal spiking Deep Neural Network (DNN) implemented on an FPGA platform. The hardware DNN system, called n-Minitaur, demonstrates a 4-fold improvement in computational speed over the previous DNN FPGA system. The proposed system directly interfaces two different event-based sensors: a Dynamic Vision Sensor (DVS) and a Dynamic Audio Sensor (DAS). The DNN for this bimodal hardware system is trained on the MNIST digit dataset and a set of unique audio tones for each digit. When tested on the spikes produced by each sensor alone, the classification accuracy is around 70% for DVS spikes generated in response to displayed MNIST images, and 60% for DAS spikes generated in response to noisy tones. The accuracy increases to 98% when spikes from both modalities are provided simultaneously. In addition, the system shows a fast latency response of only 5ms.

DOI: https://doi.org/10.1109/ISCAS.2016.7539099

Event-Driven Deep Neural Network Hardware System for Sensor Fusion

Ilya Kiselev, Daniel Neil, and Shih-Chii Liu
Institute of Neuroinformatics
University of Zurich and ETH Zurich
8057 Zurich, Switzerland

***Abstract*—This paper presents a real-time multi-modal spiking Deep Neural Network (DNN) implemented on an FPGA platform. The hardware DNN system, called n-Minitaur, demonstrates a 4-fold improvement in computational speed over the previous DNN FPGA system. The proposed system directly interfaces two different event-based sensors: a Dynamic Vision Sensor (DVS) and a Dynamic Audio Sensor (DAS). The DNN for this bimodal hardware system is trained on the MNIST digit dataset and a set of unique audio tones for each digit. When tested on the spikes produced by each sensor alone, the classification accuracy is around 70% for DVS spikes generated in response to displayed MNIST images, and 60% for DAS spikes generated in response to noisy tones. The accuracy increases to 98% when spikes from both modalities are provided simultaneously. In addition, the system shows a fast latency response of only 5ms.**

Keywords—Spiking Deep Networks, Dynamic Vision Sensor, event-driven sensors, sensor fusion, hardware spiking network

I. INTRODUCTION

Hardware systems implementing spiking Deep Networks such as Convolutional Neural Networks, Restricted Boltzmann Machines, and Deep Belief Networks (DBNs) have been demonstrated on hardware platforms including FPGAs [1][2], SpiNNaker [3], and TrueNorth. Some of these systems have been tested on recordings from event-driven Dynamic Vision Sensor (DVS) [4] to demonstrate the advantages of event-driven computing.

Hardware spiking vision systems interfaced directly to a DVS have been demonstrated for real-time applications. Although there are a few systems that are interfaced to a spiking cochlea sensor such as the Dynamic Audio Sensor (DAS) [5] and to both visual and auditory modalities [6][7], no system so far had used a spiking Deep Neural Network (DNN) in combination with both modalities.

This paper describes a real-time multi-modal DNN hardware system interfaced to two event-driven sensor modalities. A spiking DNN was previously implemented on Minitaur, an event-driven FPGA-based (Spartan 6) spiking neural accelerator system. With this system, one can implement a spiking deep network which achieves 19 million postsynaptic currents per second [6] and supports up to 65 K neurons per board. We describe the improvements made on the Minitaur architecture so that the new system called n-Minitaur is able to receive spikes in real-time from up to three event-based sensors. We conducted experiments using spiking DNNs implemented on this platform and demonstrated the classification accuracy on the MNIST dataset based on 1) the inputs from the two modalities separately and 2) the fusion of inputs from both modalities. The setup is described in Section II followed by descriptions of two experiments using this platform for visual and visual-auditory tasks in Section III and a discussion in Section IV.

This work is partially funded by the SNF project #200020-153565 "Fast Separation of Auditory Sounds" and EU HOR2020 COCOHA #644732.

II. METHODS

We first describe the DNN hardware setup, the two spiking network architectures, and the input stimuli used in this work.

A. Setup

The system consists of a Spartan-6 FPGA board and an auxiliary board with ports for direct interfacing to three spike-based sensors (Fig. 1). In this work, a DVS [4] and a DAS [5] are connected to two of these ports. The sensors communicate with n-Minitaur using an asynchronous Address Event Representation (AER) protocol. The DVS retina, with a resolution of 128x128 pixels, produces events (spike addresses) only if a pixel senses local brightness changes. The pixels output ON and OFF events which code both positive and negative changes in log-intensities respectively.

The DAS PCB [5] holds two microphones, a custom AEREAR2 binaural cochlea chip, and digital chips to handle the communication between the cochlea and the PC. The board also has an external AER interface. Each cochlea on the binaural AEREAR2 chip is modeled by a 64-stage cascaded second-order filter bank followed by a half-wave rectifier which models the inner hair cell and an integrate-fire neuron model which models the spiral ganglion cells.

B. Networks

Two DNNs are trained on the MNIST dataset following the training algorithm described in [6]. The first visual network (DNN1) of size 784-500-500-10 is trained on the 60,000 digit training set from the MNIST database (shown in Fig. 2a). The second multi-modal network (DNN2) has additional 100 audio input neurons connected to the associative layer (Fig. 2b). These neurons are driven by the spikes of the DAS. The 64

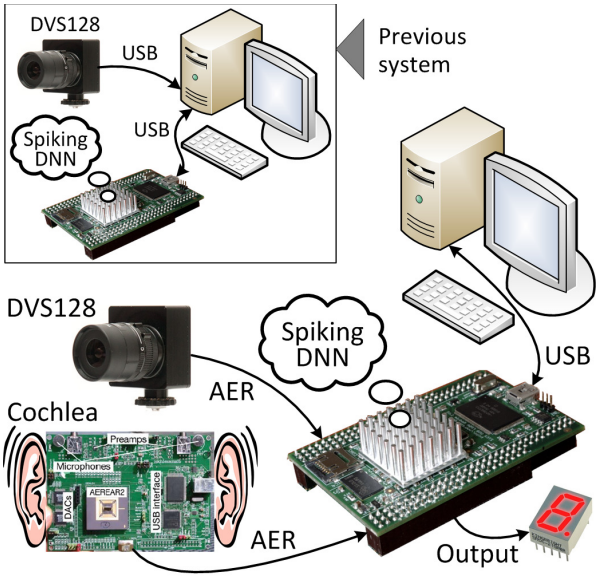

Fig. 1. n-Minitaur system interfaced directly to both DVS and DAS sensors.

cochlea channels of the DAS are mapped to the 100 neurons in the audio layer following the mapping scheme in [6]. Both networks are trained as DBNs, and then converted to feed-forward spiking neural networks (SNNs) for the classification task. The Leaky Integrate-and-Fire neuron model in the spiking DNN has a membrane potential decay time of 17.2 ms and a refractory period of 0.8 ms, as it was found to give the best performance for a practical range of the input event rates.

C. *Input Stimuli*

We describe next the two sets of test stimuli used in this work. The first set consists of artificial spike trains generated from the visual MNIST test dataset of 10000 digits. To convert the static digit images to events, spikes were generated with a probability proportional to the intensity of the pixel. These spikes were streamed to the hardware DNN using a PC.

The second set consists of spikes streamed directly from the DVS and DAS sensors without a PC in the loop. The goal of this setup is to have a complete spiking DNN hardware system which performs the classification in real-time from the sensor spikes streamed directly to n-Minitaur.

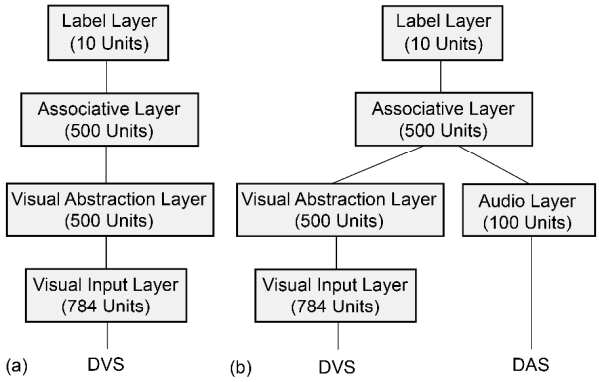

Fig. 2. Network architecture of (a) the visual DNN1 and (b) the multimodal DNN2.

TABLE I. TONE–DIGIT PAIRS IN MULTY-SENSORY FUSION TASK

Freq. (Hz)	352	371	385	414	444	470	552	588	670	720
Digit	0	1	2	3	4	5	6	7	8	9

For the visual spikes, the DVS is placed in front of an LCD display with an LED backlight. The digits are presented on the screen at a frame rate of 30 Hz. Each digit is presented for 8 times with each presentation alternating with a black screen.

For the audio spikes, a specific audio frequency is assigned to each digit similarly to [6], but with slightly different frequency values for each digit for simpler on-FPGA computation (Table I). Both the audio tone and the image of a digit are presented simultaneously in the sensory fusion experiments, and the output spikes from both sensors are recorded at the same time via USB. The classification label is determined by the neuron with the most spikes for each combination of tone and video digit.

In the real-time experiments described in Section III, the hardware sensor spike rate to n-Minitaur is reduced by dropping N-1 of N input spikes, where N ranges from 1 to 16. Both video and audio inputs are decimated using N = 10 and N = 4, respectively, to achieve an acceptable event rate of approximately 6.5 kEvents/s.

Because the performance of the DNN2 from just the DAS spikes is already close to 100% accuracy (see Table IV 5th row), an additional frequency is added to the audio stimuli in order to reduce the classification accuracy of the network. First, the frequency for each digit d is selected randomly from a normal distribution with a mean at the central frequency $F_c(d)$ and a standard deviation of 1.7. Then, a second interfering frequency is chosen from a uniform distribution ranging from 300 to 800 Hz except for a region of $F_c(d)$ ± 50Hz. The two frequencies when combined together produce an audio signal with a signal-to-noise ratio of 6dB. This set of audio stimuli corresponding to each digit produces significant error in the classification (see Table IV row #7 in Section III).

III. RESULTS

This section describes the specifications of the improved n-Minitaur implementation and the classification experiments using n-Minitaur.

A. *Improved Minitaur Architecture*

Minitaur is based on the low-cost Xilinx Spartan-6 platform. The full implementation is done on a ZTEX USB 1.15 board, which holds 128 MB of DDR2 RAM, a microSD card slot for storage, 128-kB flash memory for a bootloader, and an FX2 chip for USB interfacing. The Minitaur architecture in [1] has been improved for lower system latency and the necessity for handling AER events from up to three event-driven sensors. In order to process events arriving from multiple sensors, we modify the interface block in Minitaur which previously only received spikes from the DVS through the PC. This new architecture is shown in Fig. 3, and its specifications are shown in Table II in comparison with the original Minitaur.

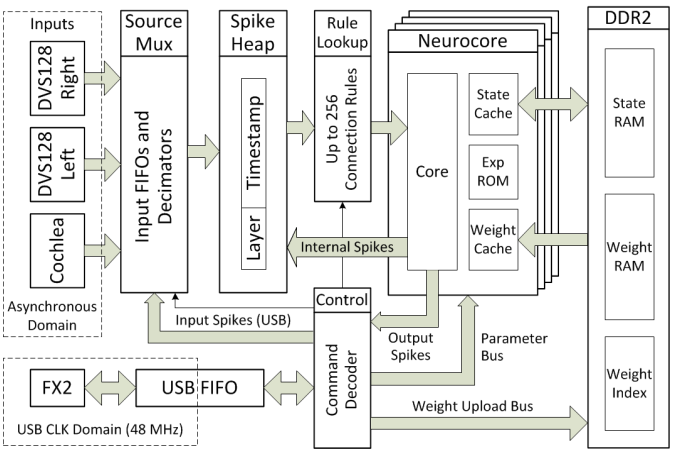

Fig. 3. New architecture of Minitaur (n-Minitaur).

The most significant changes are made on the USB interface and Control blocks, where the number of the USB-clock driven logic is reduced to a minimum. The three main improvements are described next. First, a pair of Xilinx Dual-Clock FIFOs is used to separate the clock domains.

Second, the state machines for the Spike Heap and Neurocore blocks and the weight caching strategy are redesigned for a 2.5-fold improvement of performance at the same frequencies (Table III, 2nd row).

TABLE II. SPECIFICATIONS OF MINITAUR AND N-MINITAUR

Parameter	Original (Minitaur)	Modified (n-Minitaur)
Minitaur parameters		
Number of Neurocores	32	32
Max number of neurons	65536	65536
Max number of synapses	16.78 Million	33.5 Million
Max number of connection rules	128	256
Design statistics		
Nets	173500	101255
Slice registers	17425	11669
Slice LUTs	23778	15137
Occupied slices	8845	5201
DSP blocks per Neurocore	2	1
Timing specifications		
Operating frequencies (Core/RAM)	75/132 MHz	132/264 MHz
USB download bandwidth	7.4 MB/s	30.3 MB/s
100 events download time	116 us	26 us
Min input to output (I-O) latency	293 us	238 us
I-O latency on MNIST DNN (mode)	9.2 ms	2 ms
Serial processing speed (updates/s)	0.59 Mupd/s	1.67 Mupd/s
Peak processing speed (updates/s)	18.9 Mupd/s	53.5 Mupd/s
Input spike-rate on MNIST DNN	1.95kSpikes/s	8.6 kSpikes/s

TABLE III. PERFORMANCE OF N-MINITAUR WITH DNN1 ON THE MNIST DATASET (10000 DIGITS)

Firmware	Clock (MHz)		Operation Time (s)	Time per Digit (s)	Accuracy (%)
	Core	*RAM*			
Original	75	132	5390	0.539	92.0
Modified (orig. freq)	75	132	**2084**	0.208	92.06
Modified (stable result)	105	264	**1372**	0.137	92.08
Modified (extreme)	132	264	**1359**	0.136	92.04

Third, by introducing the registered Parameter Bus, the design complexity and congestion level are reduced, and the core frequency is increased from 75 MHz to 105 MHz, leading to a 4-fold overall increase of performance.

Table III (2nd row) shows the architectural performance improvement of n-Minitaur at the same operating frequency as Minitaur. The new architecture also allows the increase of the operating frequencies of Neurocores and external RAM to 105 MHz and 264 MHz, respectively, leading to a 1.5-fold performance improvement. Further increase of the core frequency up to 132 MHz does not give much gain of performance due to the external memory interface limitations, and in addition, leads to increased routing time.

B. Hardware Inputs

Although Minitaur is capable of processing events streamed separately from event-driven visual and audio sensors using the USB interface, it was impossible to combine events from different sensors in real-time because the USB stack introduced unpredictable delays of hundreds of μs while events are captured with 1 μs precision. Moreover, data are packetized for USB transaction, so enough events need to be collected to initialize data transfer. This introduces additional delays.

On the other hand, the hardware AER interface allowed one to stream events from different AER sources into the processing module according to the timestamps of the events.

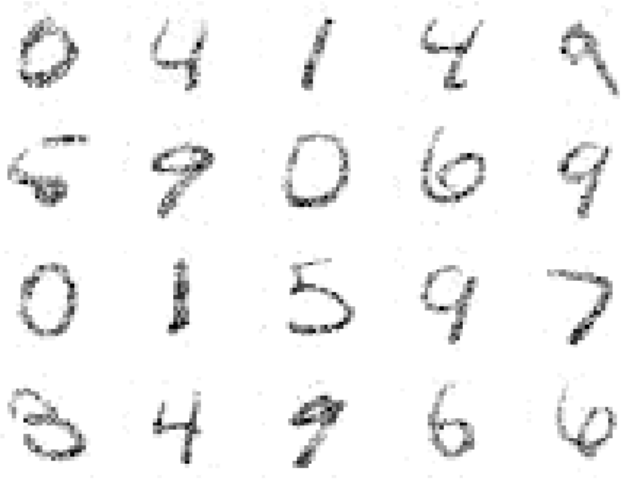

Fig. 4. Spike histograms recorded from the first layer of the network connected to the DVS via the hardware AER interface. Digits are presented to DVS as described in Section II C. 1000 events are recorded for each digit.

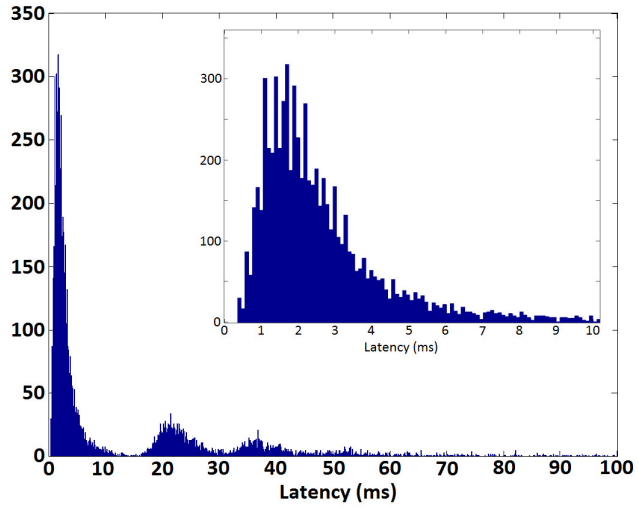

Fig. 5. Histogram of latencies from the first input spike to the first output spike based on the 10000 test digits of the MNIST dataset.

To ensure that input spikes arrive at the network in the correct order, spikes from the first layer of the network are recorded and displayed (Fig. 4). The images show that the spikes are transmitted properly to the FPGA system.

C. *Single Sensor and Sensor Fusion Experiments*

The hardware implementation of DNN1 on n-Minitaur is first tested on artificially generated spike-trains from the MNIST database and led to an accuracy of 94.1% (Table IV, Expt #1) comparable with a software simulation of the same network [6]. This accuracy decreased by 4.6% (Table IV, Expt #2), when tested with DVS spikes streamed directly into n-Minitaur through the AER interface. This decrease can be explained by the difference in the spike statistics between the artificially generated spikes and real DVS spikes.

The multi-modal network DNN2 is first tested on visual spikes using either the artificial visual spikes or the DVS spikes (Table IV, Expts #3, #4). The accuracy of the DNN2 is much lower than that of the DNN1 in this case. But the accuracy of the DNN2 using the audio input alone (Table IV, Expts #5, #6) is around 99.9%, suggesting that the DNN2 network learned to rely more on the audio input. Our hypothesis is that it was partially due to the over-simplified audio stimuli (pure tones) used to represent the digits.

TABLE IV. ACCURACY ON THE MNIST DATASET AND AUDIO TONES

#	Experiment Description	% of correctly labelled digits
1	DNN1, visual only, artificial MNIST spikes	94.1
2	DNN1, DVS spikes	89.9
3	DNN2, visual only, artificial MNIST spikes	81.5
4	DNN2, visual only, DVS spikes	70.0
5	DNN2, audio only, artificial spikes	99.9
6	DNN2, audio only, cochlea spikes	99.0
7	DNN2, audio only, cochlea + noise	60.0
8	DNN2, DVS + (cochlea + noise)	98.0

By adding audio noise, the accuracy of the network with audio-only input decreased to 60% (Table IV, Expt #7). However, by fusing spikes from the visual and noisy audio sensor inputs, the accuracy was restored to 98% (Table IV, Expt #8) , demonstrating that sensory fusion can help improve the performance of a system which performs poorly with only a single modality.

IV. DISCUSSION

This work aims to create a complete spiking hardware DNN system capable of processing spikes from event-driven audio and visual sensors. This embedded hardware system can be interfaced to a maximum of 3 AER sensors without a PC in the loop. The performance of the new DNN FPGA architecture (n-Minitaur) has been improved by a factor of 4 along with a reduction in resource utilization of 35%. The performance of the system in classifying the visual MNIST digits is of similar accuracy to that of Minitaur [1].

The hardware network shows performance increase using input events from two different event-based sensors. The measured latency between the 1st input spike and the 1st output spike is only about 5 ms (Fig. 5) when tested on the MNIST digit spikes showing the real-time operation of this system.

ACKNOWLEDGMENT

We acknowledge iniLabs and the Sensors group at the Institute of Neuroinformatics, University of Zurich/ETH Zurich.

REFERENCES

[1] D. Neil and S-C. Liu, “Minitaur, an event-driven FPGA-based spiking network accelerator,” IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. PP, no. 99, pp. 1–1, 2014.

[2] C. Zamarreño-Ramos, A. Linares-Barranco, T. Serrano-Gotarredona, and B. Linares-Barranco, "Multicasting mesh AER: A scalable assembly approach for reconfigurable neuromorphic sructured AER systems. application to ConvNets," IEEE Trans. on Biomedical Circuits and Systems, vol. 7, no. 1, pp. 82-102, Feb 2013.

[3] E. Stromatias, D. Neil, M. Pfeiffer, F. Galluppi, S. Furber, and S-C. Liu, “Robustness of spiking Deep Belief Networks to noise and reduced bit precision of neuro-inspired hardware platforms,” Front. Neurosci., doi: 10.3389/fnins.2015.00222, 2015

[4] P. Lichtsteiner, C. Posch, and T. Delbruck, “A 128 x 128 120 dB 15 µs latency asynchronous temporal contrast vision sensor,” IEEE J. Solid-State Circuits, vol. 43, no. 2, pp. 566–576, 2008.

[5] S-C. Liu, A. van Schaik, B. A. Minch, and T. Delbruck, “Asynchronous binaural spatial audition sensor with 2x64x4 channel output,” IEEE Trans Biomed. Circuits Syst., vol. 8, no. 4, pp. 453 – 464, 2013.

[6] P. O’Connor, D. Neil, S-C. Liu, T. Delbruck, and M. Pfeiffer, “Real-time classification and sensor fusion with a spiking deep belief network,” Front. Neurosci., 7:178, doi: 10.3389/fnins.2013.00178, 2013.

[7] V. Chan, C. T. Jin, and A. van Schaik, “Neuromorphic audio-visual sensor fusion on a sound-localizing robot,” Front Neurosci, vol. 6, pp. 1–9, 2012.